(4)

BBN Systems and Technologies Corporation

A Subsidiary of Bolt Beranek and Newman Inc.

Report No. 6912

COMBINING MULTIPLE KNOWLEDGE SOURCES FOR SPEECH RECOGNITION

Richard Schwartz
John Makhoul

Annual Report

Contract No. N00039-85-C-0423
September 1988

Prepared by:

BBN Systems and Technologies Corporation
10 Moulton Street
Cambridge, Massachusetts 02238

Prepared for:

DARPA
1400 Wilson Blvd.
Arlington, VA 22209

SPAWAR
2511 Jefferson Davies Hgwy.
Washington, D.C. 20363

Sponsored by
Defense Advanced Research Projects Agency (DoD)
ARPA Order No. 5425
Monitored by SPAWAR
Under Contract No. N00039-85-C-0423

COMBINING MULTIPLE KNOWLEDGE SOURCES FOR SPEECH RECOGNITION

Richard Schwartz
John Makhoul

Annual Report

Contract No. N00039-85-C-0423
September 1988

Prepared by:

BBN Systems and Technologies Corporation
10 Moulton Street
Cambridge, Massachusetts 02238

Prepared for:

DARPA
1400 Wilson Blvd.
Arlington, VA 22209

SPAWAR
2511 Jefferson Davies Hgwy.
Washington, D.C. 20363

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: BBN Report No. 6912	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Combining Multiple Knowledge Sources for Speech Recognition	5. TYPE OF REPORT & PERIOD COVERED: Annual Report 29 May 87 - 28 May 88
6. PERFORMING ORG. REPORT NUMBER	
7. AUTHOR(s): Richard Schwartz, John Makhoul	8. CONTRACT OR GRANT NUMBER(s): N00039-85-C-0423
9. PERFORMING ORGANIZATION NAME AND ADDRESS: BBN Systems and Technologies Corporation, 10 Moulton Street, Cambridge, MA 02238	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS: DARPA, 1400 Wilson Blvd., Arlington, VA 22209	12. REPORT DATE: 15 September 1988
	13. NUMBER OF PAGES: 21
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report)
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution of the document is unlimited. It may be released to the Clearinghouse, Dept. of Commerce, for sale to the general public.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Continuous Speech Recognition, Statistical Language Modelling, Search Strategies, and Multiple-Pass Search.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This annual report summarizes the work on using multiple knowledge sources in a speech recognition system. Included is research in search strategies, statistical language modeling and phonological rules. We also describe the testing of the system using the standard DARPA Resource Management Database. Additional topics of demonstrations, database documentation and porting of the system from the Symbolics Lisp machine to the SUN4 in C are also discussed.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 68 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Table of Contents

1. Introduction

This annual report covers the work performed under Contract No. N00039-85-C-0423 for Combining Multiple Knowledge Sources in Speech Recognition during the year ending May 28, 1988. The goal of this effort is to develop and refine algorithms for coordinating several sources of knowledge to perform high accuracy speech recognition in a complex military task domain with a large vocabulary, and to demonstrate the effectiveness of the developed algorithms. The application chosen for this work is the battle management task domain, in particular, a subset of the Fleet Command Center Battle Management Program (FCCBMP) application domain.

In the past year, significant progress has been made and a complete speech recognition system has been demonstrated in the FCCBMP domain with a 1000-word vocabulary, thus completing a milestone of the contract. This report gives a summary of the technical accomplishments, presented under five headings: research topics, system testing, demonstrations, database documentation, and porting the system from the LISP machine to the SUN environment. Detailed descriptions of the work are contained in three papers, which have been included in this report.

2. Research Topics

2.1 Multiple-Pass Search Strategies

An important problem in automatic speech recognition is to be able to use several diverse knowledge sources to aid in recognition. As we have stated in the past, the strategy for maximizing recognition accuracy is to consider every possible sequence of words, scoring each sequence using all relevant knowledge sources, and then to choose the sequence with the highest score or probability, given all the evidence. In practice, since the number of possible word strings is extremely large, we use several search strategies such as dynamic programming and a beam search to reduce the computation dramatically with no measurable loss in accuracy. Even so, the number of alternatives that must be considered may still be too large.

We have developed a new class of recognition search strategies, which we call multiple-pass search strategies, that will prove useful for speeding up the search with large grammars, such as large statistical grammars as well as natural language grammars. These algorithms find upper- bound scores for each of the words in the vocabulary in different regions of the input. Then, while performing grammar-directed acoustic searches, the recognizer considers only those words that are known to be likely, given the input speech. We have already demonstrated the ability of these algorithms to speed up the search with different types of grammars, including large finite-state networks, statistical grammars, and recursive transition network grammars.

The particular search strategy that we implemented is called the "Forward-Backward Search Strategy", because the first pass consists of a forward pass that computes the scores of each word ending at each possible frame. The second, or grammar pass is run backwards, using the result of the forward pass score for the words. It can be shown that this particular algorithm results in word scores that comprise a very good predictor of whether a particular hypothesis should be followed. In practice, we have found that this strategy often speeds up the computation by at least a factor of 10. In many cases, since more computation typically requires more memory, it makes the difference between being able to do the computation within the memory constraints of the machine and not being able to do it. While this particular forward-backward search does not allow maintenance of strict real-time, since part of the computation starts only after the sentence has been completed, it may make "near-real- time" possible. In addition, the resulting speed-up will be very useful in accelerating the research.

2.2 Statistical Language Modeling

In the interest of developing more robust language models to use in our speech recognition system, we have been developing a statistical language modeling technique that can be used profitably when relatively little training data is available. Typically, very large amounts of training scripts (millions of words) are required to estimate the probabilities of a statistical (Markov) language model. For applications such as the DARPA resource management task domain though, we don't expect to have more than a few thousand words of sample text for language model development purposes. Therefore, to ameliorate the estimation problem precipitated by the lack of large amounts of data, the language modeling technique we have developed estimates the probabilities of word classes rather than specific words. Thus, we use linguistic knowledge to reduce the number of probabilities that must be estimated. For example, by assuming that all names of ships are equally likely at any point in the sentence, we need only estimate the probability of the class of ships as a whole rather than the probability of each ship. Using this technique we developed a statistical language model from the training data of the DARPA 1000-word database and tested our recognition system with that grammar. The results using the statistical language model were compared with the performance of other models. When the patterns corresponding to the test sentences were included in the training for the statistical language model, the average word error was reduced by a factor of 3 relative to the Word-Pair Grammar. When the test sentence patterns were removed from the training, the performance was still approximately the same as with the Word-Pair grammar (which was trained on all the patterns). The statistical grammar is preferable, however, because it is more robust than the word-pair grammar because the former allows all possible word sequences, while the latter does not. Thus, the statistical language model -- even when trained on a small corpus of example sentences -- provides a robust grammar for new sentences.

2.3 Dialect-Dependent Phonological Rules

In our testing of the BBN BYBLOS system on the DARPA 1000- word resource management database recorded at TI, we had noticed that the recognition results for one of the speakers (RKM) (who had a southern black dialect) were significantly worse than the other speakers tested. In an effort to see whether the inclusion of dialect-dependent phonological rules would help, we constructed phonological rules specifically for this speaker. Retesting of RKM using these rules did not improve the recognition results. We concluded that, at least for this case, the inclusion of dialect-specific phonological rules does not help performance of our system.

3. System Testing

In this chapter, we summarize the various tests performed on our continuous speech recognition system, BYBLOS, and the word recognition accuracy obtained.

3.1 Speaker-Dependent Performance

Using BYBLOS, we processed the speech of eight speakers from the 1000-word DARPA resource management speaker- dependent database. Speaker-dependent models were generated for each of the speakers using 570 of their training sentences. Recognition experiments were run using the remaining 30 training sentences to verify that the models were valid. The system was then tested with an independent test set comprising 25 test sentences from each of the eight speakers. For each speaker we ran the test under two different grammar conditions: Full Branching Grammar (Perplexity = 990), and Word Pair Grammar (Perplexity = 60). With the Full Branching Grammar, the word error rate ranged from about 25% to 40% with an average of 32%; with the Word-Pair Grammar, the word error rate ranged from about 3% to 16%, with an average of 7.5%.

3.2 Live Test

On July 27, 1987, three non-BBN speakers (AS, DP, TD) who were to provide speech for the September 1987 "live tests" came to BBN to record training speech so that we can estimate speaker-dependent models for them. Each speaker read sentences during a total elapsed time of one hour, performed in two half-hour sessions. Afterwards, we listened to all of the files and deleted those sentences where the words spoken were different from those in the text transcriptions. On the average, we kept about 80% of the utterances, resulting in over 300 training utterances for each speaker or about 15 minutes of actual speech.

On September 29, the three speakers returned to test the system. The word models for each of the speakers were transfered to the Butterfly (TM) parallel processor which performed the recognition. The grammar used was the Word-Pair Grammar. Each of the speakers read 30 test sentences, one by one, and waited for the recognition answer to be typed out. All input data and recognition results were also saved on files for later analysis. On average, the recognition required about 10 times real time. This means that each sentence required about 10-40 seconds

of elapsed time. In each case, the speaker was able to finish the entire session (including putting on the microphone, comments, adjusting levels, and false starts) within 30 minutes. The word recognition error rates for the three speakers were: AS: 4.4%, DP: 5%, TD: 12%.

4. Demonstrations

BBN hosted the DARPA Speech Recognition Meeting during 13-15 October, 1987. In the workshop we demonstrated our BYBLOS continuous speech recognition system and made technical presentations on our work. The demonstrations included a near-real-time demonstration of the speech recognition being performed on the Butterfly Parallel Processor, as well as a feasibility demonstration of a complete spoken language system, in which the output of the recognizer was used to operate a simple resource management system that included the basic graphics and database operations.

5. Documentation for NBS

During the previous year we had specified the list of sentences that were used by Texas Instruments to record the DARPA 1000-word Resource Management Database, which has been sent to NBS for general distribution. During this past year we supplied NBS with documentation on the set of sentences and a complete specification of three grammars to be used for testing speech recognition systems that use this databse: a grammar of sentence patterns, a word-pair grammar that allows all word pairs that can occur in the sentence patterns, and a null grammar for the 1000 words. We defined a data format for the grammars and wrote a clear definition of test-set perplexity to be used by the community. We have assisted several DARPA sites in specifying experiments to run on their systems so that results can be compared. In addition, we have assisted CMU technically in developing their speaker-independent hidden Markov model system, and we provided Lincoln Laboratory with our phonetic dictionary.

6. Port of Software from the LISP Machine to the SUN

Because of the compute-intensive aspects of many of our new algorithms, it became very difficult to perform research to improve the performance of our system using our existing Symbolics LISP machines. During this last year we decided that we needed to change our computing environment to one that afforded sufficient computational power. After considering several alternatives, we decided that the SUN4 workstation provided a substantial increase in speed over the Symbolics machine. Therefore, we began a systematic effort at converting all of our recognition programs from LISP to C to be run on the SUN 4 workstation. Because of the change in programming language, all programs needed to be redesigned and recoded. In addition, we have designed into our programs the flexibility to include many of the variations that we expect will be tested during the coming year or two of our research.

As of May 28, we have completed the implementation of the speech decoder (recognizer) on the SUN4 workstation. Also, a large part of the training algorithm has been completed. The results of these programs are being verified by running each algorithm with the same data on the Symbolics LISP machines and the SUN4 and requiring that both the answers and the recognition scores are identical.

Our initial measurements of the speed of the new reimplemented programs has shown that we have achieved the speed advantages that we had hoped for. Specifically, the decoder, running in floating point, runs about 4 times faster than on the Symbolics machine. We coded the decoder in such a way that merely changing a compile-time flag would change whether the algorithm was performed using floating point probabilities or integer log-probabilities. When we used the latter, we achieved another factor of three increase in speed, making the decoder about 12 times faster than we had previously. The measurements of the trainer indicate that it is about 20 times faster than on the Symbolics machines. We anticipate that these large increases in speed will have a substantial impact on the amount of research that we can accomplish in the coming year.

7. Papers Presented

Details of our work have been included in three papers that were presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, April 1988, New York.

1. "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition", by P.J. Price, W.M. Fisher, J. Bernstein, and D.S. Pallett.

2. "Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database", by F. Kubala, Y. Chow, A. Derr, M.Feng, O.Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift.

3. "Statistical Language Modeling Using a Small Corpus from an Application Domain", by J.R. Rohlicek, Y.L. Chow, and S. Roucos.

All three papers are attached to this report.

Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, April 11-14, 1988, New York, N.Y. pp. 651-654

The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition

Patti Price
BBN Laboratories, Inc.
Cambridge, MA 02238

William M. Fisher
Texas Instruments, Inc.
Dallas, TX 75266

Jared Bernstein
SRI International
Menlo Park, CA 94025

David S. Pallett
National Bureau of Standards,
Gaithersburg, MD 20899

ABSTRACT

A database of continuous read speech has been designed and recorded within the DARPA Strategic Computing Speech Recognition Program. The data is intended for use in designing and evaluating algorithms for speaker-independent, speaker-adaptive speech, and speaker-dependent speech recognition. The data consists of read sentences appropriate to a naval resource management task built around existing interactive database and graphics programs. The 1000-word task vocabulary is intended to be logically complete and habitable. The database, which represents over 21,000 recorded utterances from 160 talkers with a variety of dialects, includes a partition of sentences and talkers for training and for testing purposes.

1 Introduction

The development of robust, reliable speech recognition systems depends on the availability of realistic, well-designed databases; the technical and commercial community can benefit greatly when different systems are evaluated with reference to the same benchmark material. The DARPA 1000-word resource management database was designed to provide such benchmark materials: it consists of consistent but unconfounded training and test materials that sample a realistic and habitable task domain, and cover a broad range of speakers. The goal of this database collection effort was to yield a set of data to promote the development of useful large-vocabulary, continuous speech recognition algorithms. We hope that this description will serve both to publicize the existence of the database and its availability for use in benchmark tests, and to describe the methods used in its construction.

The database includes materials appropriate to a naval resource management task. The 1000 vocabulary items and 2800 resource management sentences are based on interviews with naval personnel familiar with an existing test-bed database and accompanying software to access and display information. 160 subjects, representing a wide variety of US dialects, read sentence materials including 2 "dialect sentences" (i.e., sentences that contained many known dialect markers), 10 "rapid adaptation sentences" (designed to cover a variety of phonetic contexts), 2800 "resource management" sentences and 600 "spell-mode" phrases (words spoken and then spelled). The database is divided into a speaker-independent part and a speaker-dependent part; both are divided into training and test portions. The test portions are further divided into equal sub-parts for initial testing during system development ("development test"), and later evaluation ("evaluation test").

The methods build on and extend work by Leonard [3], Fisher *et al.* [2] and Bernstein, Kahn and Poza [1]. Original contributions of the current work include methods for designing the vocabulary and sentence set, speaker selection; and distribution of sentence material among the speakers.

The database design and implementation included: specification of a realistic and reasonable task domain, selection of a habitable 1000-word vocabulary, construction of sentences to represent the syntax, semantics, and phonology of the task, selection of a dialectally diverse set of subjects, assignment of subjects to sentences, recording of the subjects reading the sentences, and implementation of a system for the distribution and use of the database. These tasks are described in more detail below.

2 Task Design

2.1 Task Domain Specification

We chose a database query task because it is a natural place to use speech recognition technology as a human-machine interface. To define realistic constraints, and allow for eventual demonstrations of this technology, we based the task on the use of an existing, unclassified test-bed database and an interactive graphics program. The chosen task has the additional advantage that it has been the basis of much research and development in the natural language understanding community. The value of speech recognition technology is enhanced by its integration with a natural language understanding component.

The current phase of the DARPA speech recognition program specifies a 1000-word vocabulary. The test-bed database, however, has a substantially larger vocabulary size, and therefore had to be restricted. Our philosophy in selecting a 1000-word subset was to limit the number of database fields, rather than to limit the ways a user might access the information. The fields selected include information about various types of ships and associated properties: locations, propulsion types, fuel, sizes, fleet identifications, schedules, speeds, equipment availability and status. The interactive graphics commands include various ways of displaying maps and ship locations.

An initial set of 1200 resource management sentences came from: (1) preliminary interviews with naval personnel familiar with the test-bed database and the software for accessing it, and (2) systematic coverage of the database fields, subject to review by the naval personnel in follow-up interviews. These sentences were intended to provide wide coverage of the syntactic and semantic attributes of expected sentences, rather than expected relative frequencies of such sentences. Sentences were not filtered on the basis of "grammaticality", and therefore include, for example, instances of the deletion, lack of number agreement

between subject and verb, and many cases of ellipsis (i.e, omission of words required for strict grammaticality but not for comprehension, as in the deletion of the second instance of *speed* in *Is the Kirk's speed greater than the Ajax's speed*.

2.2 Vocabulary

The vocabulary was determined by collecting all words in the 1200 initial resource management sentences. If eventual users are expected to stay within the defined vocabulary, it should be, in some sense, grammatically, logically and semantically complete. Therefore, words were added so that the vocabulary included: (1) both singular and plural forms of nouns, (2) words required for all cardinal numbers less than a million, (3) words required for all ordinals needed for dates, (4) infinitive, present and past participle verb forms, (5) all months and days of the week. In addition, items were added for semantic "completeness". For example, since *high* occurred, *low*, *higher*, *highest*, *lower*, and *lowest* were added. The vocabulary was then completed by adding enough open class items to cover 33 ports, 26 other land locations, 26 bodies of water, and 100 ship names (in both nominative and possessive forms).

Since these sentences were to be read by naive subjects not familiar with the task domain or the database, the vocabulary was revised: some open class items were replaced with others thought to be easier to pronounce (*Sea of Japan* for *Sea of Okhotsk*), and spellings of some technical terms were changed to clarify the pronunciation (*TASSEM* for the acronym *TASM*).

2.3 Sentence Materials

The 1200 initial resource management sentences had some disadvantages: they included many slight variations of the same sentence (e.g., only a ship name changed or *the* deleted), and the vocabulary items were not evenly represented (the naval personnel interviewed tended to use only one or two ship names, for example, in all their examples). Further, we felt that far more than 1200 sentences would be needed to represent the vocabulary items and phonetic contexts of the task. Therefore, the initial 1200 sentences were reduced to a set of 950 unique surface semantic-syntactic patterns that were then used to generate 2800 sentences with excellent coverage of the vocabulary items.

The replacements included the replacement of instances of specific ship names with the variable [*shipname*], and of many instances of *the* with the variable [*optthe*] (to indicate optional *the*). About 300 such variables (indicated here by square brackets to distinguish them from vocabulary items) were defined and and used to replace specific instances.

In the two following examples, included to give an idea of the degree of abstraction involved, the variable definitions are: [*what-is*] ⇒ *what is, what's*; [*shipname's*] ⇒ *Kirk's, Fox's*, etc.; [*gross-average*] ⇒ *gross, average*; [*long-metric*] ⇒ *long, metric*; [*show-list*] ⇒ *show, list, show me*, etc.; [*ships*] ⇒ *carriers, cruisers*, etc.; [*water-place*] ⇒ *Indian Ocean, Sea of Japan*, etc.; [*date*] ⇒ *March 4th, 2 June 1987*, etc.

1. [*what-is*] [*optthe*] [*shipname's*] [*gross-average*] *displacement*
2. [*show-list*] [*optthe*] [*ships*] *in* [*water-place*] [*date*]

After replacement of instances with variables in the 1200 sentences, duplicates were removed, yielding 950 sentence patterns. The patterns were ordered such that those with the most unique words or classes appeared first in the list.

The 950 sentence patterns generated 2800 sentences in three passes of substitution of an instance for each variable. A counter associated with each variable determined which instance should be used for each substitution. The patterns thus generated a set of sentences that systematically covered the vocabulary items. After removal of duplicates, there were 2835 sentences. The 35 longest sentences were removed; the remaining 2800 were hand edited to remove infelicities that could arise from the procedure (such as *one carriers* generated from [*cardinal*] [*ships*]). The first 600 sentences generated were designated **training** sentences; the ordering of the patterns and the generation procedure resulted in good coverage of the vocabulary: these 600 sentences cover 97% of the vocabulary items.

In between the concept of speaker-independence (requiring no new data from new speakers) and speaker-dependence (requiring a great deal of data from each new speaker) is the concept of speaker-adaptation (requiring a small amount of data from each new speaker). For use in speaker-adaptation technologies we have provided 10 "rapid adaptation" sentences, designed to provide a broad and representative sample of the speaker's production of phonemes and phoneme sequences of the 2800 resource management sentences. The goal was to provide embedded sets of one, two, five and ten sentences that each had the best coverage (for its size) of the relevant phonemic material. Thus, the first is the best adaptation sentence, the second sentence, when added to the first, is the best combination of two sentences according to the same coverage criteria, and so on up to ten.

A coverage score was calculated for each phoneme and phoneme pair in a sentence based on the observed frequency of the phoneme or phoneme pair in the 2800 sentences, but breadth of coverage was promoted by dividing the observed frequency of each phoneme or phoneme pair by a factor (we used 3.0) each time it was used in the material currently having a score calculated. In order to inhibit the tendency for the longest (and most difficult to read) sentences from being selected, we normalized by dividing the score by sentence length. The resulting adaptation sentences are listed in the appendix.

For the "spell-mode" utterances, 600 words were selected from the 1000 vocabulary items; the 400 words not selected were inflected variants of those chosen.

3 Subject Selection and Recording

3.1 Subject Selection

On the basis of demographic and phonetic characteristics, 160 subjects were selected from a set of 630 adults who had participated in an earlier database effort [2]. These 630 native speakers of English (70% male, 30% female) with no apparent speech problems formed a relatively balanced geographic sample of the United States. As a group, the subjects were young, well-educated, and White; 63% in their twenties, 78% with a bachelors degree and 4% Black. Each speaker was identified with one of eight geographic regions of origin: New England, New York, Northern, North Midland, South Midland, Southern, Western, or 'Army Brat' (people who moved around a lot while growing up).

Among other material, each of these 630 subjects had recorded two dialect-shibboleth sentences (i.e., sentences containing several instances of words regarded as a criterion for distinguishing members of dialect groups). These sentences, included in the appendix, were hand-transcribed and used to derive a phonetic profile of each speaker as to phonology, voice quality, and manner of speaking. The 630 speakers were automatically divided into 20 clusters according to their pronunciation of several consonants, speaking rate, F0, and phonation quality. From these 630 speakers (now identified by phonetic cluster, geographic origin and demographic characteristics) 160 were selected for the speaker-independent part of the database, and 12 for the speaker-dependent part.

The 160 speaker-independent subjects were chosen to satisfy the following constraints, in order: 1) even distribution of subjects over four geographic regions (NE-NY, Midland, South, North-West-or-Army) and over the 20 clusters derived from observed phonetic characteristics; 2) 70% male, 30% female. These constraints are satisfied in the subject selection, and each major division of the database (training, development test and evaluation test) have similar distributions across sex and geographic origin.

The 12 speaker-dependent subjects were chosen to satisfy the following constraints: 1) representation of each of the 12 largest phonetic clusters; 2) seven male, five female; and 3) geographical representation as follows: one each from New York and New England, and two each from Northern, North Midland, South Midland, Southern, and Western. Of the 12 selected speakers, 11 were from the speaker-independent part of the database, and all were relatively fluent readers with no obvious speech problems.

3.2 Subject-Sentence Assignment

Both the speaker-independent and speaker-dependent parts of the database are divided into sets for training, development test and evaluation test.

In the speaker-independent training part of the database, 80 speakers each read 57 sentences (40 resource management sentences, the 2 dialect sentences, and 15 spell-mode phrases). 1600 distinct resource management sentences were covered in this part of the database; any given sentence was recorded by two subjects. The distribution of sentences to speakers was arbitrary, except that no sentence was read twice by the same subject. Each of the 80 speakers read 15 spell-mode phrases, yielding 1200 productions covering 300 unique words. Each spell mode phrase in this part was read by 4 speakers.

In the speaker-independent development test set and evaluation test set, 40 speakers each read 30 resource management sentences, the 2 dialect sentences, the 10 rapid adaptation sentences, and 15 spell-mode phrases. 600 resource management sentences were randomly selected for each test and assigned to the 1200 available productions (40 speakers times 30 sentences), yielding two productions per sentence, as in the training phase. Similarly, in each test set, 150 spell-mode phrases were selected and assigned to the 600 available spell-mode productions.

The following table illustrates the structure of the speaker-independent part of the database. The numbers indicate how many sentences each subject read. The total number of resource management sentences covered by each subset of the database is indicated in parentheses. These are referred to as "types" in the table in distinction to sentence tokens, or productions by a particular speaker. In all, for the speaker-independent database, 9120 sentences were recorded (4560 for training, 2280 for development test, and 2280 for evaluation test). Note that, this being the speaker-independent database portion, the training subjects do not overlap with those in the test parts of the database.

SPEAKER-INDEPENDENT DATABASE

	training		development test		evaluation test	
No. Subjects	80		40		40	
No. Sentences (types)						
Resource Management	40	(1600)	30	(600)	30	(600)
Dialect	2	(2)	2	(2)	2	(2)
Adaptation	0	(0)	10	(10)	10	(10)
Spell-mode	15	(300)	15	(150)	15	(150)
TOTALS	57	(1902)	57	(762)	57	(762)

For the speaker-dependent training portion of the database, each of 12 subjects read the 600 resource management training sentences, the 2 dialect sentences, the 10 rapid adaptation sentences, and a selection of 100 spell-mode phrases. The 1200 spell-mode readings covered 300 word types, with 4 productions per word.

In the speaker-dependent test portion of the database, these same 12 speakers each read 100 resource management sentences for the development-test part of the database and another 100 resource management sentences for the evaluation-test part, as well as 50 spell-mode phrases. From the 2200 resource management sentences not read in the training phase, two random selections of 600 sentences were made, one for the development test and one for the evaluation test portion. Distributing these over the productions available in each gives 2 utterances per sentence. Similarly, two random selections of 150 words each were made from the pool of 600 spell-mode phrases for the development and evaluation test sets. Distributing these over the 600 readings available yields 4 productions per word.

The following table illustrates the structure of the speaker-dependent part of the database. Again, the total number of different resource management sentences ("types") covered in each subset is indicated in parentheses after the number indicating how many sentences were read by each subject. In all, for the speaker-dependent database, 12,144 utterances were recorded (8544 for training, 1800 for development test, and 1800 for evaluation test). As is appropriate for a speaker-dependent database, the speakers in the training set are the same as the speakers in the test set.

SPEAKER-DEPENDENT DATABASE

	training		development test		evaluation test	
No. Subjects	12		12		12	
No. Sentences (types)						
Resource Management	600	(600)	100	(600)	100	(600)
Dialect	2	(2)	0	(0)	0	(0)
Adaptation	10	(10)	0	(0)	0	(0)
Spell-mode	100	(300)	50	(150)	50	(150)
TOTALS	712	(912)	150	(750)	150	(750)

3.3 Recording Procedure

The utterances were digitally recorded in a sound-isolated recording booth on two tracks: one from a Sennheiser HMD414 headset noise-cancelling microphone, and the other from a B&K 4165 one-half inch pressure microphone positioned 30 cm from the subject's lips, off-center at a 20 degree angle. The material was digitized at 20,000 16-bit samples per second per channel, and then down-sampled to 16,000 kHz.

Prompts appeared in double-high letters on a screen for the subject to read. After the recording, both the subject and the director of the recording session listened to the utterances and re-recorded those with detected errors. Any pronunciation considered normal by the subject was accepted.

4 Database Availability and Use

This database, which is intended for use in designing and evaluating algorithms for speech recognition, is being made available to provide: (1) a carefully structured research resource, and (2) benchmarks for performance evaluation to judge both incremental progress and relative performance.

At present only the data from the Sennheiser microphone is available. This material alone amounts to approximately 930 Megabytes (MB) of data for the speaker-dependent subset and 640 MB for the speaker-independent subset, with an additional 460 MB included in the spell-mode subset. The down-sampled (16 kHz) data in Unix "tar" format (6250 bpi) can be made available on a loan, copy and return basis.

To provide benchmark test facilities, a set of procedures and a uniform scoring software package have been developed at the National Bureau of Standards (NBS). The scoring software implements a dynamic programming string alignment on the orthographic representations for the reference sentences and for the system outputs. Comparable scoring necessitated agreement on a standard orthographic representation for each vocabulary item. The scoring software and testing procedure are being used in the DARPA program for performance evaluation, and are available to the general public on request [4].

For those organizations wishing to determine and report performance data corresponding to that reported by DARPA program participants, NBS can provide test material used in DARPA benchmark tests [4]. If the results are to be publicly reported, it is required that the summary statistics be obtained using the NBS scoring software, and that copies of system output for these tests be made available to NBS.

5 Conclusion

For DARPA program participants, this database has proven useful in the design and evaluation of speaker-independent, speaker-adaptive, and speaker-dependent speech recognition technologies; we hope it will be useful to others as well. Similarly, the methods developed for its design and collection should prove useful in the development of similar databases.

We have described the characteristics of the DARPA 1000-word resource management database: the task domain, the vocabulary, the sentence materials, the subjects, the division into training and testing portions. We have also described the steps involved in creating this database, including the recording procedure and new methods for designing the vocabulary and sentence set, speaker selection, and distribution of sentence materials among the speakers. In addition, we have outlined procedures for obtaining the database and for using it as a benchmark. Further details on each of these areas will be made available with the database.

Acknowledgements. This effort has been a collaborative effort that involved not just the authors, but the DARPA speech recognition community in general. However, chief responsibility for the various tasks required by the project was assigned as follows: BBN - task design, vocabulary selection and sentence construction; SRI - subject selection and dialect sentences; TI - subject-sentence assignment and recording of data; NBS - distribution and evaluation methods. We gratefully acknowledge the naval experts who helped us and the DARPA Strategic Computing Speech Recognition Program for funding this effort (contract numbers N00039-85-C-0423, N00039-85-C-0338 and N00039-85-C0302 monitored by SPAWAR, and, for NBS, DARPA order number 6079).

References

[1] Bernstein, J., M. Kahn and T. Poza (1985) "Speaker sampling for enhanced diversity," *IEEE ICASSP-85*, paper 41.2.

[2] Fisher, W., V. Zue, J. Bernstein and D. Pallett (1987) "An acoustic-phonetic database," *J. Acoust. Soc. Am., Vol.81, Suppl. 1*, abstract OO1.

[3] Leonard, R. G. (1984) "A database for speaker-independent digit recognition," *IEEE ICASSP-84*, paper 42.11.

[4] For further information on availability of the database, test procedures and scoring software, contact D. S. Pallett, Room A216 Technology Building, National Bureau of Standards, Gaithersburg, MD, 20899. Telephone: (301) 975-2935.

APPENDIX

Dialect-Shibboleth Sentences

1. She had your dark suit in greasy wash water all year
2. Don't ask me to carry an oily rag like that.

Rapid Adaptation Sentences

1. Show locations and C-ratings for all deployed subs that were in their home ports April 5.
2. List the cruisers in Persian Sea that have casualty reports earlier than Jarrett's oldest one.
3. Display posits for the hooked track with chart switches set to their default values.
4. What is England's estimated time of arrival at Townsville?
5. How many ships were in Galveston May 3rd?
6. Draw a chart centered around Fox using stereographic projection.
7. How many long tons is the average displacement of ships in Bering Strait?
8. What vessel wasn't downgraded on training readiness during July?
9. Show the same display increasing letter size to the maximum value.
10. Is Puffer's remaining fuel sufficient to arrive in port at the present speed?

Presented at the IEEE International Conference and Acoustics, Speech and Signal Processing, April 11-14, 1988, New York, N.Y. pp. 291-294.

Continuous Speech Recognition Results of the BYBLOS System on the DARPA 1000-Word Resource Management Database

F. Kubala, Y. Chow, A. Derr, M. Feng*, O. Kimball, J. Makhoul.
P. Price, J. Rohlicek, S. Roucos, R. Schwartz, and J. Vandegrift

BBN Laboratories Incorporated, Cambridge, Ma. 02238

*Northeastern University, Boston, Ma. 02115

ABSTRACT

We present results of the BBN BYBLOS continuous speech recognition system tested on the DARPA 1000-word resource management database. The system was trained in a speaker dependent mode on 28 minutes of speech from each of 8 speakers, and was tested on independent test material for each speaker. The system was tested with three artificial grammars spanning a broad perplexity range. The average performance of the system measured in percent word error was: 1.4% for a pattern grammar of perplexity 9, 7.5% for a word-pair grammar of perplexity 62, and 32.4% for a null grammar of perplexity 1000.

1 INTRODUCTION

A meaningful comparison between the performance of speech recognition algorithms and systems can be made only if the systems have been tested on a common database. Even with common testing material, comparative results become difficult to interpret when grammars are used to constrain the recognition search. The ambiguity introduced by the use of grammars can be overcome by reporting results with the grammar disabled, which would establish a baseline acoustic recognition performance for the system, and by using standard generally available grammars. Finally, reporting a standard measure of the constraint provided by a grammar makes the results more meaningful.

In this paper we report results for the BBN BYBLOS system tested on a standard database using two well defined, artificial grammars and with an unconstrained null grammar. The database has been developed by the DARPA Strategic Computing Speech Recognition Program for the purpose of comparative system performance evaluation of continuous speech recognition systems [6].

In section 2, we describe the BYBLOS system. In section 3, the database and testing protocol are discussed. The grammars used in the experiments are described in section 4. Section 5 presents the recognition system results. The results are discussed in section 6.

2 THE BYBLOS SYSTEM

The BYBLOS continuous speech recognition system [2] uses discrete density hidden Markov models (HMM) of phonemes, a phonetic dictionary, and a finite state grammar to achieve high recognition performance for language models of intermediate complexity. The parameters of the HMMs are estimated automatically from a set of supervised training data. The trained phoneme models are combined into models for each word in the dictionary. These phonetic word models are then used to compute the most likely sequence of words in an unknown utterance. A formal description of a complete HMM system is presented in [1].

The BYBLOS system has been designed to accomodate large vocabulary applications. It trains a set of phoneme models which requires only a moderate amount of speech to adequately observe all the phonemes. In addition, the system trains a separate model for each distinct context in which a phoneme is observed. A phoneme's context can be defined by its adjacent phonemes or the word in which it appears. Context modeling captures coarticulation phenomena explicitly and preserves phonetic detail for those contexts which occur frequently in the training material [7]. By combining the smoothed phoneme models with the detailed context models, BYBLOS makes maximal use of the available training material. The performance improvement gained by using context dependent phoneme modeling has been reported in [3].

After training is completed, the dictionary is populated by compiling the trained phonetic models into word networks. A finite state grammar, if used, is compiled from a formal language model specification. To decode

an unknown utterance, BYBLOS utilizes the precompiled knowledge sources jointly in a time-synchronous, top-down search. This search strategy allows efficient pruning and minimizes local decisions.

BYBLOS has been demonstrated in a speaker dependent and a speaker adaptive mode. Speaker dependent modeling achieves high performance by estimating the model parameters from a training corpus which is large enough to contain most of the contexts likely to appear in subsequent use of the system. The speaker dependent mode has been used to achieve the results reported in this paper. The speaker adaptive mode modifies the well trained, speaker dependent word models of one speaker to model a new speaker. This technique allows the system to benefit from the well trained word models of a prototype speaker even when the training material from the new speaker is extremely limited. The adaptation mode of the BYBLOS system is discussed in [4,8].

3 Database

The database, described in detail in [6], was designed to provide a standard for research in speaker dependent, speaker adaptive, and speaker independent continuous speech recognition. The database was designed to cover the vocabulary, syntax, and functionality of a naval resource management task. The vocabulary consists of 1000 words. The task domain covered by the database is specified by a set of 950 sentence patterns which were used to generate the 2800 distinct sentences in the database.

The speaker dependent database provides 600 sentences (about thirty minutes of speech) designated as training material from each of twelve dialectally diverse speakers, collected in six different sessions. The scripts for the training material are designed to maximize coverage of the vocabulary and sentence patterns. The speakers include seven male and five female speakers. Independent test material was collected for the twelve speakers during additional sessions.

The experiments reported in this paper have been conducted for the purpose of comparative performance evaluation within the DARPA community. The evaluation was administered by the National Bureau of Standards (NBS). For the speaker dependent portion of the evaluation, tests were conducted using eight of the twelve available speakers.

We withheld 30 sentences from the training material for each speaker to be used for adjusting global system parameters. The remaining 570 sentences that we used for training include 952 unique words from the vocabulary. Approximately 5% of the words in the dictionary are not observed at all in the training set, 36% occur only once, and 49% occur more than once.

Twenty five sentences were selected by NBS as test material for each speaker. The test sets are different for each speaker, but on average, each set contains about 200 words. The test sentences for the eight speakers cover 46% of the dictionary. 91% of the word tokens occurring in the eight test sets have occurred more than once in the training set illustrating the effectiveness of the training data coverage over the task domain.

4 Grammars

The results reported below have been run using three different grammar conditions. These grammars are not intended as serious models of the task domain, but are used because they are simply defined and allow the system to be tested over a broad range of language model constraint.

A straight-forward measure of the constraint provided by a grammar is *test set perplexity* [5] which is measured on a finite state network generated by the grammar and a given set of test sentences. For the purpose of perplexity measurement, a distinguished symbol designating inter-sentence silence is added to the dictionary and to the end of each sentence of the test set. The augmented sentences are then concatenated and appended to an initial inter-sentence silence to form the word sequence, $w_1, w_2, \ldots, w_n$. If the word sequence is sufficiently long, the probability of the sequence given the grammar, $\hat{P}(w_1, w_2, \ldots, w_n)$, can be used to compute an estimate of the grammar perplexity.

The perplexity of the grammar, given the test set word sequence, is defined as:

$$L = 2^K \tag{1}$$

where

$$K = -\left(\frac{1}{n}\right)\sum_{i=2}^{n} \log_2 \hat{P}(w_i | w_{i-1}, \ldots, w_1) \tag{2}$$

is the average per word entropy of the language model, and

$$\hat{P}(w_1) = 1 \tag{3}$$

For the grammars used in these experiments, the probabilities on the words allowed by the grammar at position i in the test set word sequence are assumed to be uniform.

The three grammars, which we call the sentence pattern, word-pair, and null grammar, allow all sentences in the training and test databases. The sentence pattern grammar is compiled directly from the set of 950 sentence patterns covering all sentence types in the task domain [6]. The perplexity of the pattern grammar, averaged over the eight speakers' test sets, is 9. The word-pair grammar allows all two-word sequences allowed in the sentence pattern gram-

mar. Its perplexity is about 62. The null grammar allows all sequences of words in the vocabulary and therefore offers no language model constraint. The effective perplexity of the null grammar is equal to 1000 — the vocabulary size.

5 Results

The system parameters for these experiments were derived from two speakers' data collected at BBN and limited testing on two speakers from the DARPA database (CMR and BEF) using the data that we withheld from the training set. The system configuration was then fixed for the entire set of experiments. Each speaker was trained only once.

The database speech was collected at Texas Instruments (TI) in a sound isolating booth. For these experiments we used speech sampled at 20 kHz, through a Sennheiser HMD-414, close-talking, noise-canceling microphone. 14 Mel-scale-warped cepstral coefficients were computed every 10 ms. using a 20 ms data window, and vector quantized using an 8-bit codebook.

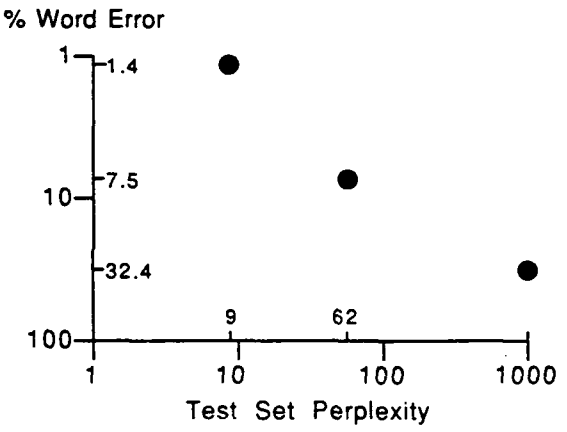

Figure 1: Recognition Performance as a Function of Grammar Perplexity. The axes are log scale.

Figure 1 shows recognition performance, averaged across the eight speakers, for the three grammar conditions. The performance is given in percent word error:

$$\text{WORD ERROR} = 100 \times (S - D - I) / N$$

where:

S = number of substitution errors,

D = number of deletion errors,

I = number of insertion errors,

N = total number of word tokens in the test sentences.

This measure has been proposed as a standard within the DARPA community. Note that since the number of insertion errors possible is not bounded, this error measure can exceed 100%.

A word hypothesis is counted in error if it does not identically match the correct word transcription. Specifically, homophones (e.g., to, two, too; or ships, ship's, ships') are counted as errors. Homophone errors typically occur only in the null grammar experiments where they account for approximately 4% of the word error rate. Furthermore, no special significance is given to errors which are phonetically close to the correct answer (minimal pair differences) or to errors which leave the semantic interpretation of the sentence intact (most deletions of the word 'the').

Individual results for each speaker are shown in Table 1. Two speakers, CMR and DTD, are female. The results are given as word error, defined above, and as word correct:

$$\text{WORD CORRECT} = 100 \times [1 - (S + D) / N]$$

where, S, D, and N are defined as before.

Note that:

$$\text{WORD ERROR} \neq 100 - \text{WORD CORRECT}.$$

For the pattern and word-pair grammars, the sentence error rate and test set perplexity are also given. For the null grammar case, the sentence error rate is near 90%, and the perplexity = 1000.

6 Discussion

In our experience, average word error (E) for a set of speakers can be estimated as a function of perplexity (L) by:

$$E = \alpha\sqrt{L} \tag{4}$$

Figure 1 indicates that $\alpha \approx 1$ for this data set over most of the perplexity range. We have conducted numerous experiments on speech collected at BBN in normal office environments. The experiments have used a variety of grammars including those reported here. We consistently find the average word error to be reasonably predicted by using $\alpha = \frac{1}{2}$ which is half the error rate obtained for the TI speakers. The difference in average performance between the TI and BBN data may be explained by differences in speaking style and rate. The speakers collected at BBN have some experience with speech recognition systems and generally speak more clearly than the speakers collected at TI.

While the average performance is generally predicted by perplexity, an individual speaker's performance may not be. For example, speaker DTB performs far below average for the null grammar but above average for the word-pair and pattern grammars. Similarly, the performance for RKM on the word-pair grammar is far worse than would be predicted from his results on the pattern or null grammar.

It is clear from these results that performance can be

	Sentence Pattern				Word-Pair				No Grammar	
	word error %	word correct %	sentence error %	test set perplexity	word error %	word correct %	sentence error %	test set perplexity	word error %	word correct %
BEF	2.6	98.3	20	8	8.9	93.2	44	62	40.9	62.6
CMR	2.7	99.1	20	7	9.3	94.7	52	66	39.6	65.4
DTB	0.5	100.0	4	10	5.4	96.5	32	64	39.4	63.1
DTD	1.0	99.0	8	8	6.7	94.2	44	54	26.7	75.3
JWS	0.9	99.1	8	9	4.3	96.2	28	59	25.6	75.4
PGH	0.5	99.5	4	9	6.0	96.0	24	56	32.0	70.5
RKM	2.4	98.1	16	10	16.4	89.7	52	64	30.5	71.8
TAB	0.5	100.0	4	9	3.2	97.7	20	67	24.8	76.5
avg	1.4	99.1	10.5	9	7.5	94.8	37.0	62	32.4	70.1

Table 1: Recognition Performance by Speaker for three grammar onditions.

made arbitrarily high by lowering the grammar perplexity. For large vocabulary, complex task domain applications, however, low perplexity grammars are likely to be too restrictive for real use. We expect that habitable grammars for 1000 word task domain applications will require perplexities larger than 50.

Acknowledgement

This work was supported by the Defense Advanced Research Projects Agency and was monitored by the Space and Naval Warfare Systems Command under Contract No. N00039-85-C-0423.

References

[1] Bahl, L.R., F. Jelinek, and R.L. Mercer (1983) "A Maximum Likelihood Approach to Continuous Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence* Vol. PAMI-5, No. 2, March 1983, pp. 179-190.

[2] Chow, Y., M. Dunham, O. Kimball, M. Krasner, F. Kubala, J. Makhoul, P. Price, S. Roucos, and R. Schwartz (1987) "BYBLOS: The BBN Continuous Speech Recognition System," *IEEE ICASSP-87*, paper 3.7.1.

[3] Chow, Yen-Lu, Richard Schwartz, Salim Roucos, Owen Kimball, Patti Price, Francis Kubala, Mari O. Dunham, Michael Krasner, John Makhoul (1986) "The Role of Word-Dependent Coarticulatory Effects in a Phoneme-Based Speech Recognition System," *IEEE ICASSP-86*, paper 30.9.1.

[4] Feng, M., F. Kubala, R. Schwartz (1988) "Improved Speaker Adaptation Using Text Dependent Spectral Mappings," *IEEE ICASSP-88*, Elsewhere in these proceedings.

[5] Jelinek, F. (1987) "Self-Organized Language Modeling for Speech Recognition," Unpublished manuscript, IBM T. J. Watson Research Center, Yorktown Heights, NY.

[6] Price, P., W. Fisher, J. Bernstein, and D. Pallett (1988) "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition," *IEEE ICASSP-88*, Elsewhere in these proceedings.

[7] Schwartz, R. M., Y. L. Chow, O. A. Kimball, S. Roucos, M. Krasner, and J. Makhoul (1985) "Context Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech," *IEEE ICASSP-85*, paper 31.3.

[8] Schwartz, Richard, Yen-Lu Chow, Francis Kubala (1987) "Rapid Speaker Adaptation using a Probabilistic Spectral Mapping," *IEEE ICASSP-87*, paper 15.3.1.

Presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, April 11-14, 1988, New York, N.Y. pp. 267-270

STATISTICAL LANGUAGE MODELING USING A SMALL CORPUS FROM AN APPLICATION DOMAIN [1]

Jan Robin Rohlicek *Yen-Lu Chow* *Salim Roucos*

BBN Laboratories Incorporated
Cambridge MA 02238

ABSTRACT

Statistical language models has been successfully used to improve performance of continuous speech recognition algorithms. Application of such techniques is difficult when only a small training corpus is available. This paper presents an approach for dealing with limited training available from the DARPA *resource management* domain. An initial training corpus of sentences was abstracted by replacing sentence fragments or *phrases* with variables. This training corpus of phrase sequences was used to derive parameters a Markov model. The probability of a word sequence is then decomposed into the probability of possible phrase sequences and the probabilities of the word sequences within each of the phrases.

Initial results obtained on 150 utterances from six speakers in the DARPA database indicate that this language modeling technique has potential for improved recognition performance. Furthermore, this approach provides a framework for incorporating linguistic knowledge into statistical language models.

1 INTRODUCTION

This paper addresses the use of statistical language modeling techniques in continuous speech recognition in the DARPA 1000-word *naval resource management* application domain [5]. This application involves the recognition of "natural" speech queries to an interactive database system. As will be discussed below, the "language" which will be used is unknown and a large training corpus is not available. Straightforward application of statistical language modeling techniques is therefore difficult. However, a language model is required to obtain very good recognition performance.

Language models provide a way of assigning likelihoods to word sequences in a language. The combination of such a measure with a measure of the acoustic likelihood of a word sequence has been shown to give good recognition performance in many applications. Several approaches have been successfully employed for languages of various complexity and various sizes of training corpus (for example [2]).

In certain restricted domains, finite state grammars have been used with considerable success (see [4] for example). In this case, the likelihood of a word sequence is a binary decision — a sequence is either parsed in the grammar or it is not in the allowable language. The extent to which the actual word sequences in the application are parsed by the grammar is termed *coverage*. When the language is known and not complex, the coverage is generally high and the constraints are well modeled by the grammar.

In the case of large vocabularies (> 1000 words) and "natural" language input one approach taken is the specification of formal grammars which describe the syntactic and semantic constraints of the domain [6]. The important issue is then the extent to which this grammar provides sufficient coverage while ruling out invalid word sequences. It has been found that it is difficult to achieve a high degree of coverage however. Recognition performance is generally high on sequences parsed by the grammar. However, when coverage of the valid word sequences is not high, then the language model actually introduces errors by not allowing valid word sequences.

To overcome the performance constraints imposed by poor coverage, statistical language models can be used. When a large training corpus is available, the parameters of a statistical language model can be determined. To the extent that the training corpus is representative of the real application, such techniques provide good performance [1]. Furthermore, since no binary decision as to the validity of a word sequence is necessary, the method is less "brittle" than the formal grammar techniques.

In the domain of interest in this paper, the language is not sufficiently well defined to allow the use of a finite-state grammar which both captures the constraints of the domain and is of reasonable size. Furthermore, there is no adequate training corpus for construction of a straightforward statistical model to characterize the word sequences. Due to the natural language interface, a grammar describing the complete language is very complex. Also, it is difficult to evaluate the extent to which any particular grammar

[1] This research was supported by the Defense Advanced Research Projects Agency under contract N00039-85-C-0423 monitored by SPAWAR

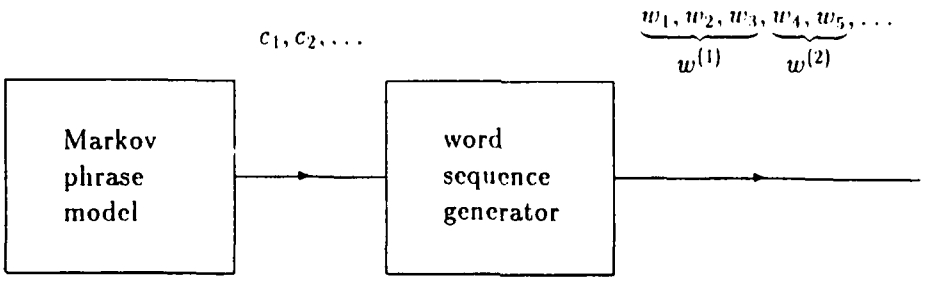

Figure 1: Word sequence model

covers sentences of the ultimate application domain. The complexity of the language suggests a statistical approach. However, since the application does not yet exist, a truly representative training corpus is not available. Furthermore, we feel that due do heavy use of jargon and unusual sentence structure, any attempt to use a training corpus from another domain, such as general English text, would be ineffective.

The approach described in this paper attempts to incorporate some linguistic knowledge of the structure of the language into a probabilistic framework. Using this approach, we will show very good performance can be obtained when the algorithm is evaluated on sentences which are independent of those used in construction of the statistical model.

In the next section, the basic structure of the model is described followed by a description of the training method employed. In Section 3, the results on six speakers from the DARPA database are presented. Finally, Section 4 contains a short discussion and concluding remarks.

2 Approach

2.1 Language Model Structure

The principle goal in the design of the probabilistic language model is to allow the estimation of robust model parameters from the modest training corpus which is available. A Markov model used to generate word sequences directly has too many parameters (the transition probabilities) to be estimated reliably from the limited training corpus. By considering a simpler model, which has fewer parameters associated with it, robust estimates might be obtainable. Furthermore, some linguistic structure can be identified, and this structure is incorporated into the model.

The model for the generation of a word sequence is composed of two part (Figure 1). First, a sequence of *phrase* variables $c_1, c_2, \ldots$, is generated as a Markov chain. Then, for each phrase c_i a sequence of words $w^{(i)}$ is generated, independent of the phrases c_j, $j \neq i$. The probability of a phrase sequence $c_1, c_2, \ldots c_N$ is

$$\Pr(c_1, \ldots, c_N) = \Pr(c_1)\Pr(c_2|c_1)\cdots\Pr(c_N|c_1, \ldots, c_{N-1})$$

The probability of the phrase sequence and the word sequence $w_1, w_2, \ldots, w_n$ is then

$$\Pr(c_1, \ldots, c_N, w_1, \ldots, w_n) = \sum_{N, w^{(1)}, \ldots, w^{(N)}} \prod_{i=1}^{N} \Pr\left(w^{(i)} \mid c_i\right) \Pr(c_i \mid c_1, \ldots, c_{i-1})$$

where the sum is effectively over the possible segmentations of the word sequence into the phrases. *Note that since any $w^{(i)}$ might be a null expansion of a phrase, this representation of the probability in fact has an infinite number of terms.*

Using this structure, we identify phrases based on syntactic and semantic components of the language. For example, typical phrases include "open" set classes such as ship names or complex expressions such as dates. Also, to complete the coverage of the language, single word phrases are also allowed. Associated with each phrase is a small finite state grammar describing all possible ways that a phrase can be expanded.

The parameters of the Markov phrase model are derived from the training corpus. The probabilities $\Pr\left(w^{(i)}|c_i\right)$ associated with the transformation of phrases into word subsequences are assigned *a priori*. In this way, a small training corpus can be used to estimated the smaller number of parameters of the Markov model without sacrificing the robustness of the overall model.

2.2 Corpus

In the resource management application domain, the initial training corpus consists of approximately 1200 sentences on a vocabulary of about 1000 words which are thought to be representative of the domain. These sentences were generated attempting to simulate the interaction of a person with the interactive database system. This database is further described in [5] in these proceedings.

From these initial sentences, a set of approximately 1000 sentence *patterns* were generated. This process was carried out manually. The goal was to incorporate linguistic knowledge by replacing syntactically and semantically similar components of the sentences with phrase identifiers. For

example, a typical sentence and its corresponding pattern is

> What gas surface ships which are in Coral Sea are SLQ-32 capable
> $\Longrightarrow$ *what* |*prop-type*| *surface* |*vessels*| |*optthat-are*| *in* |*water-place*| *are* |*capability*| *capable*

A phrase such as |*optthat-are*| can be expanded into the finite state grammar

$$\begin{aligned} |\textit{optthat-are}| &\longrightarrow \text{(empty string)} \\ &\longrightarrow \text{which are} \\ &\longrightarrow \text{that are} \end{aligned}$$

For each experiment, these patterns were partitioned into a training and testing set. The testing set was not used in the estimation of the model parameters. The test sentences were generated from the test patterns by expanding the phrases into word sequences.

2.3 Parameter Estimation

For each speaker, a set of 900 training patterns was chosen which was disjoint of the patterns of the test sentences. A first order Markov model was constructed based on the training patterns (the patterns included the context of the sentence initial and sentence final boundary markers). The transition probabilities were obtained from the relative frequencies of phrases pairs in the training patterns, using a simple interpolation rule to incorporate part of the zeroth order distribution. Interpolation is used to overcome limitations of insufficient training by assigning reasonable nonzero probabilities to all event. Specifically, if $F(c_i|c_{i-1})$ is the relative frequency of c_i following c_{i-1} and $F(c_i)$ is the relative frequency of c_i then probability of a phrase c_i is assumed to be

$$\Pr(c_i|c_1,\ldots,c_{i-1}) = \lambda F(c_i|c_{i-1}) + (1-\lambda)F(c_i)$$

where in these experiments $\lambda = 0.9$ for all states. For each grammar associated with a phrase, a simple assumption that all possible word sequences are equally likely was made. Specifically, if there are m different non-null expansions of a phrase c, then each of these expansion $w_1,\ldots w_k$ is assigned a probability

$$\Pr(w_1,\ldots,w_k|c) = (1-\theta_c)\frac{1}{m}$$

where θ_c is the probability of a null expansion. For non-optional phrases, $\theta_c = 0$.

2.4 Decoding Method

The decoding algorithm used to generate the results is based on the algorithm presented in [2,3]. A hidden Markov model approach is taken in which context-dependent triphone models are trained using the "forward-backward" algorithm. Whole word models are constructed by concatenation of interpolated (by context) triphone models.

The statistical language model described above is combined with these word models. Conceptually, each transition in the Markov phrase model is replaced by a network representation of the sub-grammar associated with the phrase (with branching probabilities at each of the nodes). Each arc in the grammar is replaced by the hidden Markov model for the word associated with the arc. Therefore, the entire model can be thought of a one large hidden Markov model.

The decoding algorithm attempts to find the maximum likelihood phrase sequences $c_1,\ldots c_N$ and the word expansions $w^{(i)}$ of each phrase. The output word sequence is then the concatenation of the $w^{(i)}$.

3 Results

Initial experiments were conducted on a speaker not included in the DARPA database in order to determine suitable system parameters (which were then unchanged).

3.1 Test on Training

Before evaluation on the independent test sets, two speakers were run using sentences derived from patterns in their training sets. As expected, the perplexity Q [2] for the statistical model is very low in this case and recognition word error rate[3] is small. As shown in Table 1 this demonstrates

	test on training			
speaker	MP	(Q)	WP	(Q)
dtb	5.4%	42.5	5.4%	(69.8)
pgh	4.5%	40.3	5.9%	(53.3)

Table 1: Word error rate on training set (MP=Markov phrase model; WP=word pair grammar)

that when evaluated on the training set such a statistical model give low perplexity and good recognition performance. For comparison, results using a grammar (WP) is shown. This grammar is constructed to allow all two-word sequences which occur in any expansion of the training patterns. Note that even though the statistical model used incorporates the interpolation rule described above, and therefore allows all possible word sequences and not simply those in the the WP grammar, the perplexity is lower

[2] Perplexity $Q = 2^I$ where I is the average information ($-\log_2 p$) of the state transitions (with probabilities p) in a set of sentences using a particular probabilistic model.

[3] Word error rate is the average number of substitution (S), deletion (D) and insertion (I) errors per reference word ($= (S + D + E)/N$ where N is the number of reference words).

than the WP grammar and the performance is somewhat better.

3.2 Test Results

The full evaluation consisted of six speakers from the DARPA database with 25 utterance each. The word error rates are presented in Table 2. In order to evaluate

	independent test		test on training
speaker	MP ($Q \approx 75$)	NG ($Q = 1000$)	WP ($Q \approx 60$)
bef	12.3%	40.9%	8.9%
cmr	13.8%	39.6%	9.3%
dtb	11.8%	39.4%	5.4%
dtd	10.0%	26.7%	6.7%
pgh	7.0%	32.0%	6.0%
tab	6.3%	24.8%	3.2%
ave.	10.2%	33.9%	6.6%

Table 2: Recognition word error rate (MP=Markov phrase model; NG=null grammar; WP=word pair grammar)

the improvement due to the statistical language model, the word error rate for a "null" grammar (NG) in which all word sequences can occur is also shown. The NG result is a measure of the acoustic difficulty of the task. The result using the word-pair (WP) grammar, trained on the training *and testing* patterns is also presented in order to show that the statistical approach achieved almost equal performance without the loss imposed by imperfect coverage. Also, note that the perplexity of the statistical model ($Q \approx 75$) is comparable to the WP grammar ($Q \approx 60$)[4] despite the fact that the WP perplexity is measured on a subset of its training sentences. Finally, in order to evaluate the effect of coverage of a grammar on overall performance, consider the sentence error rates of 49% for the statistical MP case and 36% for the WP grammar. In order for the WP grammar to achieve 49% sentence error rate including the effect imperfect coverage, at least 80% of the sentences would have to parse[5]. Currently, this level of coverage is not available.

The results presented are preliminary. Several aspects of this approach have not been investigated. For instance, the structure of the Markov model has not been fully explored. Though some experiments have been performed to evaluate the use of certain higher order states which have been observed in the training, it is not clear how the model should be constructed to actually improve recognition performance significantly. Also, the assumption that all word sequences within a grammar are equally likely is clearly a very crude approximation and some improvement may be obtainable through more careful assignment of these probabilities.

[4]Perplexity on the WP grammar is obtained assuming all branches in a deterministic finite state network representation are equally likely.

[5]Suppose a fraction of sentences x parse under the WP grammar. Assuming the remainder have a sentence error rate of 36%, then the overall error rate would be $(1 - x) + 0.36x$. For this to be less than 49%, $x > 80\%$.

4 Conclusions

The results presented here demonstrate the viability of incorporating linguistic structure into a statistical model. In the resource management domain, neither solely statistical nor linguistic techniques alone are adequate at this time. Straightforward statistical techniques lack sufficient training and linguistic techniques have an inadequate coverage. However, the combination of the modest training available and simple linguistic abstractions of this training corpus provides good performance.

References

[1] L. R. Bahl, F. Jelinek, R. L. Mercer. "A Maximum Likelihood Approach to Continuous Speech Recognition". *IEEE Trans. Pattern Analysis and Machine Intelligence*, 5(2):179-190, March 1983.

[2] Y.-L. Chow, M. O. Dunham, O. A. Kimball, M. A. Krasner, G. F. Kubala, J. Makhoul, P. J. Price, S. Roucos, R. M. Schwartz. "BYBLOS: The BBN Continuous Speech Recognition System". *IEEE ICASSP*, Dallas Texas, April 1987.

[3] F. Kubala, Y. Chow, A. Derr, M. Feng, O. Kimball, J. Makhoul, P. Price, J. Rohlicek, S. Roucos, R. Schwartz, J. Vandegrift. "Continuous Speech Recognition Results on the BYBLOS System on the DARPA 1000-Word Resource Management Database". *IEEE ICASSP*, New York, New York, March 1988.

[4] B. Lowerre, R. Reddy. "The HARPY Speech Understanding System". 1980. in *Trends in Speech Recognition*, W. A. Lea (ed.), pp. 340-360, Englewood Cliffs, N.J.: Prentice Hall.

[5] P. Price, W. Fisher, J. Bernstein, D. Pallet. "The DARPA 1000-Word Resource Management Database for Continuous Speech Recognition". *IEEE ICASSP*, New York, New York, March 1988.

[6] J. J. Wolf, W. A. Woods. "The WHIM Speech Understanding System". 1980. in *Trends in Speech Recognition*, W. A. Lea (ed.), pp. 316-339, Englewood Cliffs, N.J.: Prentice Hall.